Rudra Murthy

Ph.D. Scholar at IIT Bombay

E-mail: rudra@cse.iitb.ac.in
Mobile: +91 976 922 9057
http://murthyrudra.github.io

Objective

Effective borrowing of features from one or more assisting languages to improve performance of various NLP tasks on low-resource languages.

Educational Qualifications

Pursuing Ph.D. in Computer Science, under guidance of *Prof. Pushpak Bhattacharyya*, at Indian Institute of Technology Bombay, Mumbai.

Qualification	Board/University	Year	CPI / %
Ph.D (Compute Science and Engineering)	Indian Institute of Technology, Bombay	Since July 2013 to *present*	8.64 CPI
M.Tech (Computer Science)	Indian Institute Of Science, Bangalore	July 2013	5.9(8)
B.E. (Computer Engineering)	RNS Institute Of Technology, Bangalore	2011	81.08 %

Relevant Course Work

IISc Bangalore

1. Data Mining
2. Linear Algebra

IIT Bombay

1. Natural Language Processing
2. Foundation of Machine Learning

Research Work

Ph.D.

- **Multilingual Learning for Natural Language Processing using Deep Learning**.(*Current*)
  Ph.D Thesis Topic under guidance of Prof Pushpak Bhattcharyya.

  Description: Deep Learning techniques have become the de-facto approach for any Natural Language Processing (NLP) task. Deep neural networks coupled with unsupervised learning (in the form of pre-trained word embeddings or initial training of language model followed by supervised training) have revolutionised the area of NLP, at least for resource-rich languages. However, the success of deep learning techniques needs to be taken with a pinch of salt. The deep learning techniques have known to perform badly for low-resource languages when trained on very small data. To be fair even traditional machine learning models would perform badly if not for feature engineering. Due to limited data the model cannot reliably establish correlations with features and class labels leading to poor performance.

  In my thesis, I focus on borrowing features (implicitly statistics) from a related language (also known as multilingual learning). This should minimise the impact of data sparsity and lead to improvements in the low-resource language for the task in hand. We apply the above intuition to Named Entity recognition (NER) task. We show that borrowing features from a related languages infact helps improve the NER performance in low-resource languages. However, the improvements gained is limited by the lexical gap between the two languages. Though the lexical gap can be overcome by use of crosslingual embeddings, these embeddings have known to be of poor quality for Indian languages. Currently, my research is focused on obtaining better crosslingual embeddings for Indian languages.

Publications:

- Rudra Murthy and Pushpak Bhattacharyya, **A Deep Learning Solution to Named Entity Recognition**, *International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, Konya, Turkey, 3-9 April, 2016.
- Rudra Murthy, Anoop Kunchukuttan and Pushpak Bhattacharyya, **Judicious Selection of Training Data in Assisting Language for Multilingual Neural NER**, *Association for Computational Linguistics (ACL)*, Melbourne, Australia, 15-20 July, 2018.
- Rudra Murthy, Mitesh Khapra and Pushpak Bhattacharyya, **Improving NER Tagging Performance in Low-Resource Languages via Multilingual Learning**, *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 2019.
- Rudra Murthy, Anoop Kunchukuttan and Pushpak Bhattacharyya, **Addressing word-order Divergence in Multilingual Neural Machine Translation for extremely Low Resource Languages**, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, USA, 2-7 June, 2019.

M.Tech.

- **Learning from Positive and Unlabeled Examples** (*Aug 2012 to July 2013*)
  M.Tech dissertation under guidance of Prof. Shirish K. Shevade at IISc Bangalore.

  Objective: In many real life cases, it is easy to collect positive examples. It is difficult to define a negative set, but, instead one can collect large unlabeled data. This unlabeled data contains a mixture of both positive and negative examples. Traditional classifiers cannot be directly used in this setting and requires modification.

  We explored use of Pairwise Ranking based Logistic Regression model to the problem. The motivation for using logistic regression is to get a confidence score from the system. We hoped that this confidence score can be used for better judgement of the class label. We obtained mixed results by beating the baseline on some datasets and performing closer on some datasets.

Internship

- **Multilingual Models for Language Identification in Code-Mixed Data** (*May 2016 to July 2016*)
  Internship under guidance of Mitesh Khapra at IBM Bangalore.

  Objective: We study the performance of deep learning model for language identification task. We experiment with two training strategies. Training individual models for every language pairs involved. Training a single model (multilingual model) for all language pairs jointly. We observe multilingual models to be beneficial for improving the Named Entity tagging performance. However, this comes at a cost as we observed a drop in identification performance for some of the languages.

Technical Skills

- **Programming Language**: C/C++, Java, Python.
- **Platform**: Linux.
- **Deep Learning Framework**: Torch, PyTorch

Personal Details

Contact Address:	Hostel 1 (Room: 56) IIT Bombay, Mumbai (IN) – 400 076
E-mail:	rudra@cse.iitb.ac.in rudramurthy@.iitb.ac.in

Date: February 26, 2019
Place: IIT Bombay, Mumbai

Rudra Murthy V.